# Early pest detection in soy plantations from hyperspectral measurements: a case study for caterpillar detection

Matías Tailanián[a], Enrique Castiglioni[b],Pablo Musé[ab], Germán Fernández Flores[a], Gabriel Lema[a], Pedro Mastrángelo[a], Mónica Almansa[a], Ignacio Fernández Liñares[a] and Germán Fernández Liñares[a].

[a]CSI Ingenieros, Montevideo, Uruguay
[b] Department of Signal Processing, Universidad de la República, Montevideo, Uruguay

## ABSTRACT

Soybean producers suffer from caterpillar damage in many areas of the world. Estimated average economic losses are annually 500 million USD in Brazil, Argentina, Paraguay and Uruguay. Designing efficient pest control management using selective and targeted pesticide applications is extremely important both from economic and environmental perspectives. With that in mind, we conducted a research program during the 2013-2014 and 2014-2015 planting seasons in a 4,000 ha soybean farm, seeking to achieve early pest detection. Nowadays pest presence is evaluated using manual, labor-intensive counting methods based on sampling strategies which are time consuming and imprecise. The experiment was conducted as follows. Using manual counting methods as ground-truth, a spectrometer capturing reflectance from 400 to 1100 nm was used to measure the reflectance of soy plants.

A first conclusion, resulting from measuring the spectral response at leaves level, showed that stress was a property of plants since different leaves with different levels of damage yielded the same spectral response. Then, to assess the applicability of unsupervised classification of plants as healthy, biotic-stressed or abiotic-stressed, feature extraction and selection from leaves spectral signatures, combined with a Supported Vector Machine classifier was designed. Optimization of SVM parameters using grid search with cross-validation, along with classification evaluation by ten-folds cross-validation showed a correct classification rate of 95%, consistently on both seasons.

Controlled experiments using cages with different numbers of caterpillars - including caterpillar-free plants - were also conducted to evaluate consistency in trends of the spectral response as well as the extracted features.

**Keywords:** Soy plant, defoliation, point spectrometer, spectral signature, UAV, multispectral camera, biotic stress, abiotic stress, Support Vector Machine

## 1. INTRODUCTION

Damages produced by caterpillar in soy plantations is nowadays a major concern for producers since it incurs in significant losses, but also for environmental issues. Indeed, with the increase of soybean production, the application of pesticides and fertilisers has become a source of major impact on soil and potable water when plantations are close to rivers, since it causes the appearance of toxins and toxic algae. Recently in Uruguay, for instance, the government is paying important attention to this issue. From then on, the capability of precisely spotting the presence of caterpillars in soy plantations has become a major issue.

In this work, we propose a method to detect the presence of caterpillar in soy plants indirectly, by inferring the stress the plants are suffering as a consequence of defoliation. While the final goal is to perform this detection remotely by capturing images with an UAV, this paper is focused on field measurements using a point spectrometer. There are two fundamental reasons to follow such approach: First, the cost of hyperspectral

---

cameras is still too high, as well as drones having the capability to use them as payload; second, the acquisition of a continuum spectral reflectance is extremely useful to understand which are the main features that differentiate a plant under stress (either biotic or abiotic) from a healthy one.

All field experiments were performed using a so-called point spectrometer* This spectrometer has a field of view of $30^o$, so the captured spectral reflectance is computing by integrating over a neighbourhood of the spotted region. The reflectance spectrum captured by the instrument ranges from 340nm to 1170nm. For each wavelength $\lambda$, the spectral reflectance $\rho(\lambda)$ is measured as the ration between the reflected light and the incident light, that is

$$\rho(\lambda) = 100\frac{G_r(\lambda)}{G_i(\lambda)},$$

where $G_r(\lambda)$ and $G_i(\lambda)$ are the reflected and the incident spectral intensity, respectively. It follows that a fundamental step prior to taking measurements is to calibrate the incident light. This calibration is performed by measuring the reflectance on a white, Lambertian surface.

Our team was composed by engineers and data scientists, an agronomist with deep knowledge of the farm, and an experienced entomologyst. This allowed us to spot the regions where caterpillars were present, and to establish the ground truth.

The paper is organised as follows. In Section 2 we discuss previous work. Then, in Section 3 we describe the data acquisition protocol, both for the acquisition of the soy plants' spectral signature, as for the multispectral images capture from the drone. Section 4 is devoted to the analysis of the acquired data. It will proven that by extracting a set of features from their signatures, soy plants under stress can be automatically detected with an accuracy of 95%. A controlled experiment using cages that isolate the plants with and without caterpillars will be described in Section 6. We conclude in Section 7, where a discussion on the results of our experiments and future work are presented.

## 2. PREVIOUS WORK

Recently, in 2010, a book edited by Jones and Vaughan ("Remote sensing of vegetation"[1]) presents a large variety of technical information related to the use of hyperspectral cameras in drones for pests and diseases detection, and nutrient management. The book "Hyperspectral remote sensing of vegetation and agricultural productivity", edited by Thenkabail, Lyon and Huete in 2011,[2] is nowadays the most complete and in depth reference on the field.

There exists a myriad of successful studies to many cultures and plagues, however, to our knowledge, no previous work on the detection of caterpillar in soy plantations has been reported in the literature. In this section we briefly summarise a few previous works that we consider to be the most relevant to our investigation.

McKinion et al. (2009)[3] show that plagues in cotton, in particular Lygus lineolaris or Palisot de Beauvois, can be controlled at an early stage using NDVI followed by a clustering or a segmentation stage. The study concludes that, based on experiments performed on 1200 ha in Mississippi, USA, the application of such a system may reduce the use of pesticide in 40%.

Yang et al. (2009)[4] propose a technique to distinguish stress induced by greenbugs (Schizaphis graminum) and Russian wheat aphids en wheat based on remote sensing. The study concludes that using a 16 bands radometer or multispectral camera, it is possible to distinguish the stress caused by these two insects. The work presented by Backoulou et al. (2011[5] and 2013[6]) exhibits similar methodologies in wheat to distinguish the stress induced by Diuraphis Noxia from other stresses causes either biotic or abiotic, and to quantify the difference.

Finally, the group leaded by Susan Ustin at UC Davis has been studying the application of remote sensing with hyperspectral imaging in general. An interesting work, that dates back to 2002 (see for instance Zhang et al., 2002;[7] Zhang et al., 2003[8]) studies stress detection tomatoes, caused by late blight disease. This study shows that it is possible to identify, diagnose and classify the infection degree by sensing the tomato plants' canopy.

---

*In our case, we use a BLUE-Wave Spectrometer, from StellarNet Inc., USA.

## 3. DATA ACQUISITION CAMPAIGN

### 3.1 Collection of spectral signatures with a field spectrometer

Two kind of experiments were performed using the point spectrometer. In the first one, we spotted, respectively, a set of 79 and 92 healthy and stressed plants due to caterpillar defoliation, and we built a labeled database for each class. In each plant, three leaves with different defoliation levels were chosen, and for each one three spectrometer measurements were captured. A first observation was that, for each plant, all the spectral responses were almost identical, and therefore stress was actually a property of each plant.

The second kind of experiment was performed on a more controlled environment, using cages covered with a white tulle. Two cages were used to cover a pair healthy plant, seemingly free of any caterpillar; other two cages were used to cover healthy plants, also seemingly free of caterpillars, where we added about YY stage-ZZ caterpillars. Stress measurements were taken before setting the cages, and a couple of weeks later.

### 3.2 Capture of multispectral images from a UAV

In this experiment, three caterpillar-free plants were chosen and isolated with the same kind of cages described above. Then, a different number of caterpillars were manually introduced in each of them. A couple of weeks after the cages were a set caterpillars were introduced, we flew over the region with a drone having a camera with blue, red and near infrared channels. The goal was to see wether the defoliation stress due to the caterpillars could be seen in the NDVI images with the naked eyes or not. Note that this fact is of major concern, since in an uncontrolled situation, the producers could immediately spot those regions infected by caterpillars, and then apply pesticides only on those regions.

## 4. DATA PROCESSING AND CLASSIFICATION

The purpose of this section is to propose an automatic detection method for plants under stress *via* their spectral signature. A typical reflectance spectral responses of healthy and caterpillar stress-induced plants is shown on Figure 1. The most noticeable difference in the shape of the spectral signatures can be seen on the regions marked in red, green and black. In the red region (number 1), we can see the presence of a dip in the response



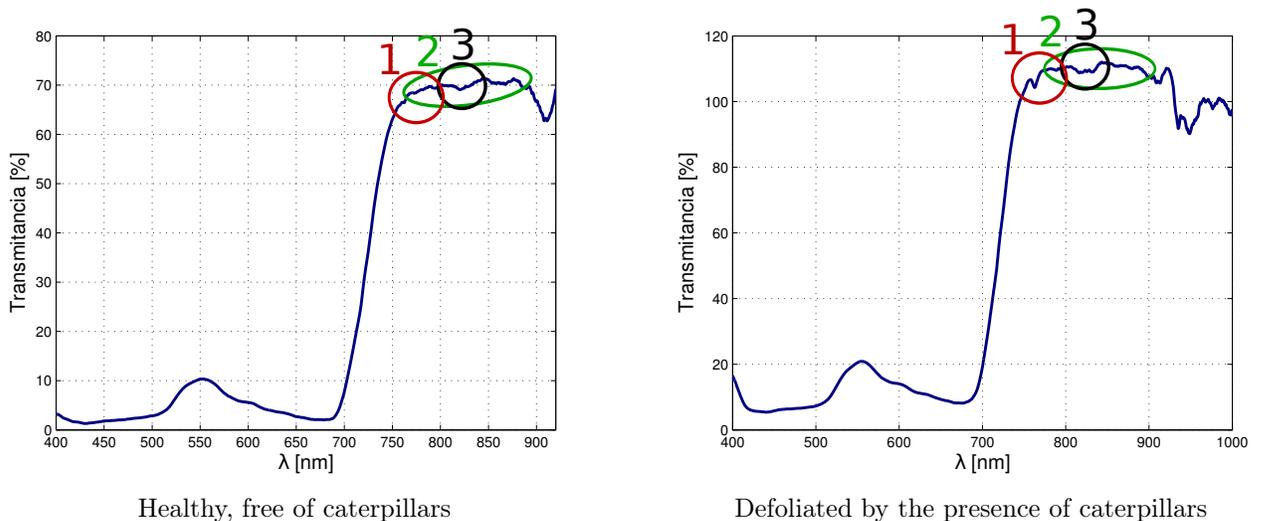Healthy, free of caterpillars        Defoliated by the presence of caterpillars

Figure 1: Typical reflectance of healthy and caterpillar defoliated plants. Note the differences, specially in regions marked as 1, 2 and 3.

of the stressed plant, that is not present in the healthy one. In the green region (number 2) we can see that the slope is significantly different between both signatures. Globally, for the healthy plants, the slope is positive, while the stressed plant exhibits zero slope or even slightly negative. Finally, for the region marked in black (number 3), the valley seems to be more pronounced than for the healty one.

### 4.0.1 Feature extraction for spectral signature classification

In order to do automatic detection of caterpillar-induced stress, we need to quantify the features described in Section 4. There is a wide variety hyperspectral vegetation indexes, and studies of their relationships with agricultural crop characteristics (see for instance[9, 10]). In the following we present the features that are extracted to design a classifier. Some of them are based on the literature, while others were specifically designed for our problem by carefully inspecting systematic differences between curves of the two classes.

### Index 760

This index is defined as

$$I_{760} = \frac{\rho_{760} - \rho_{720}}{\rho_{760} + \rho_{720}},$$

where $\rho_\lambda$ denotes reflectance at $\lambda$ nm.

### Normalized Difference Vegetation Index

This index is very popular in agriculture. It is defined as

$$NDVI = \frac{\rho(850) - \rho(670)}{\rho(850) + \rho(670)},$$

where the wavelengths are also expressed in nm.

### Derivative in the region 780nm–880nm

This index captures the slope observed in the green region (Number 2) in the previous section. In order to measure it, the curve is first smoothed to avoid localized variations. Then, the index is computed as the maximum or the mean derivative. Another possibility consist in fitting the smoothed curve with a straight line using least squares.

### Fitting the dip with a parabola

The goal is to fit the curve with the equation $ax^2 + bx + c = 0$ in the dip within the red region (number 1). The feature we consider is the quadratic term $a$. Figure 2 shows an example. It is worth mentioning that the range used to fit the parabola is the one defined by the closest inflexion points.
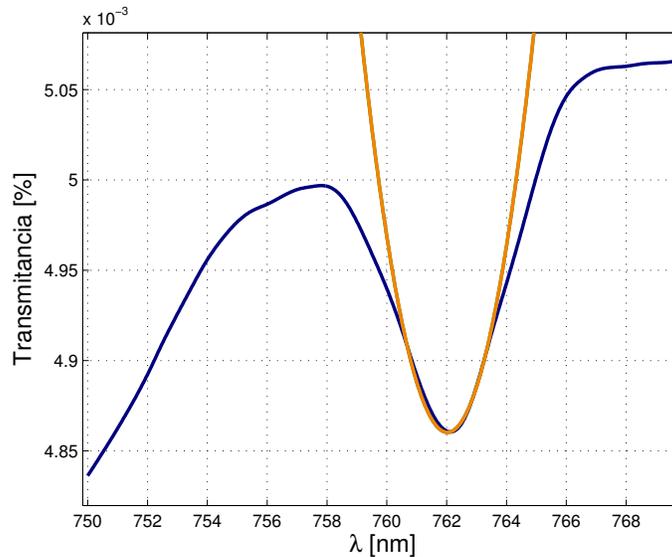


Figure 2: Parabola fitting of the dip within the red region (number 1) Figure 1.

**Dip shape factor**

This feature is, to some extent, redundant with the previous one. The idea is to measure the ration between the depth of the dip and the region where it was fitted. It is computed as the ration $x/y$, where $x$ and $y$ are defined as shown in Figure 3.
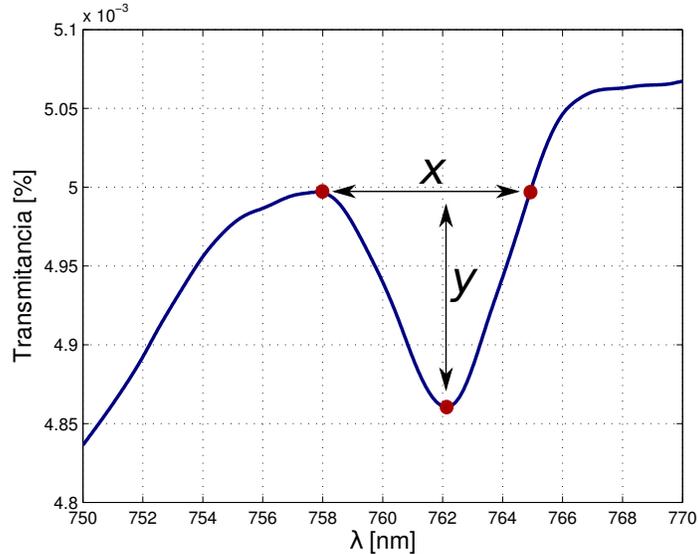


Figure 3: Factor de forma del dip

**Clusters visualisation**

In order to visualise the separation between the samples corresponding to healty and caterpillar stress-induced plants, we chose three indices among the set of indices defined above. Of course, as long as the number of samples is large enough, the larger the number of indices, the largest the is the separation. Here the three selected indices were chosen with the only goal to show that for our problem, classification seems to be well posed. Figure 4 illustrates the two set of samples in the subspace defined by (dip ratio, maximal derivative in 780–880nm, NDVI). The blue points correspond to samples of healthy plants, while the orange ones correspond to caterpillar stress-induced plants. It seems clear that both clusters are almost linearly separable. The classification results obtained in Section 5 will show that, using all features and a nonlinear boundary, a good separation can be obtained.

## 5. RESULTS: AUTOMATIC CLASSIFICATION OF PLANTS FROM SPECTRAL FEATURES

Now that we have defined a set of features that seem to be well adapted to the differences in the spectral signatures between healthy and stressed plants, we address the problem of designing a classifier following a typical machine learning approach.[11, 12] The classification method that yielded the best results was C-SVM, a support vector machine with margin.[13] An optimal choice of all parameters was performed using grid search, and features were chosen in closed loop with the classifier (the so-called *wrapper mode* methodology[14]). We train a classifier using our labeled data, and then we decompose the dataset in a training set and a validation set. To evaluate the performance of the classifier, we use a 10 folds cross validation technique.

Thus, among the 163 plants that were inspected, only 8 of them were wrongly classified. *This amounts to a more than 95% of classification success.*
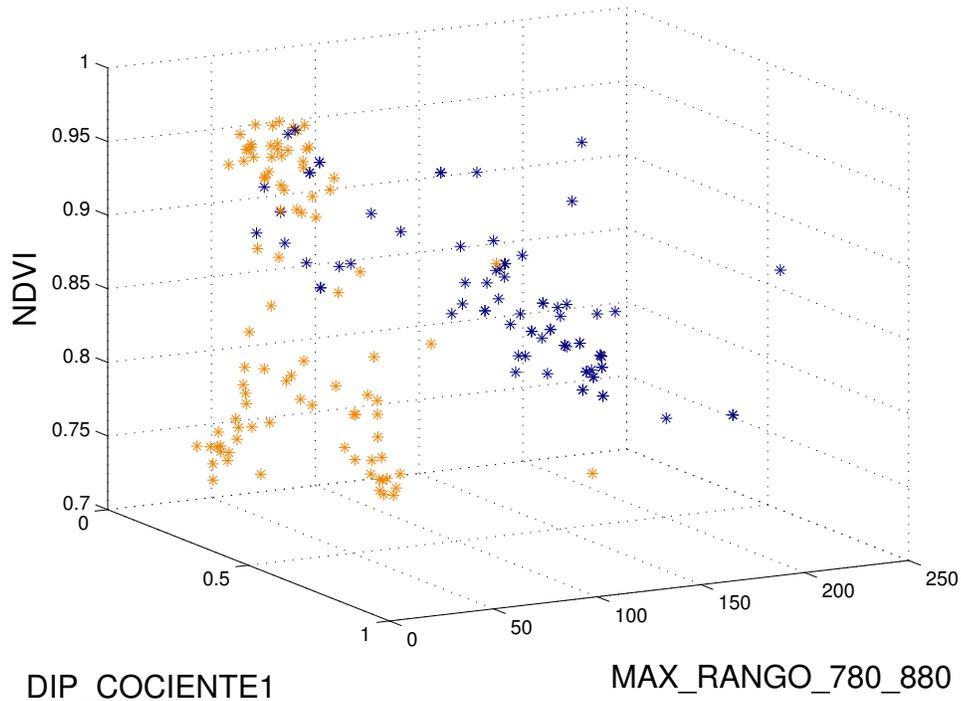
Figure 4: Visualisation of healthy and stressed plants in the subspace (dip ratio, maximal derivative in 780–880nm, NDVI). The blue points correspond to samples of healthy plants, while the orange ones correspond to caterpillar stress-induced plants.

|  | Classified as healthy | Classified as stressed |
|---|---|---|
| True healthy | 76 | 3 |
| True stressed | 5 | 87 |

Table 1: Confusion matrix for a set of 79 healthy plants and 92 caterpillar stress-induced plants.

## 6. CONTROLLED EXPERIMENTS USING CAGES

The experiment was conducted as follows. We selected four plants that exhibited very similar conditions, and we covered them with a tulle cage in order to prevent the flow of caterpillars. In two of them (*Caterpillar cage 1* and *Caterpillar cage 2*), a set of caterpillars were added. The other two cages (*Control cage 1* and *Control cage 2*) were used as control. Before closing the cages, for each of the plants, the spectral response of three leaves was acquired. Then, nine days later, the same measurements were performed. The goal of this experiments was to analyse changes in the spectral response in *Caterpillar cage 1* and *Caterpillar cage 2*, compared to changes in the control cages. Figures 5 and 6 show respectively, the spectral responses in the control cages, and their conterparts *Caterpillar cage 1* and *Caterpillar cage 2*. The curves in blue correspond to the first measurements, while the red ones correspond to the measurements acquired nine days later.
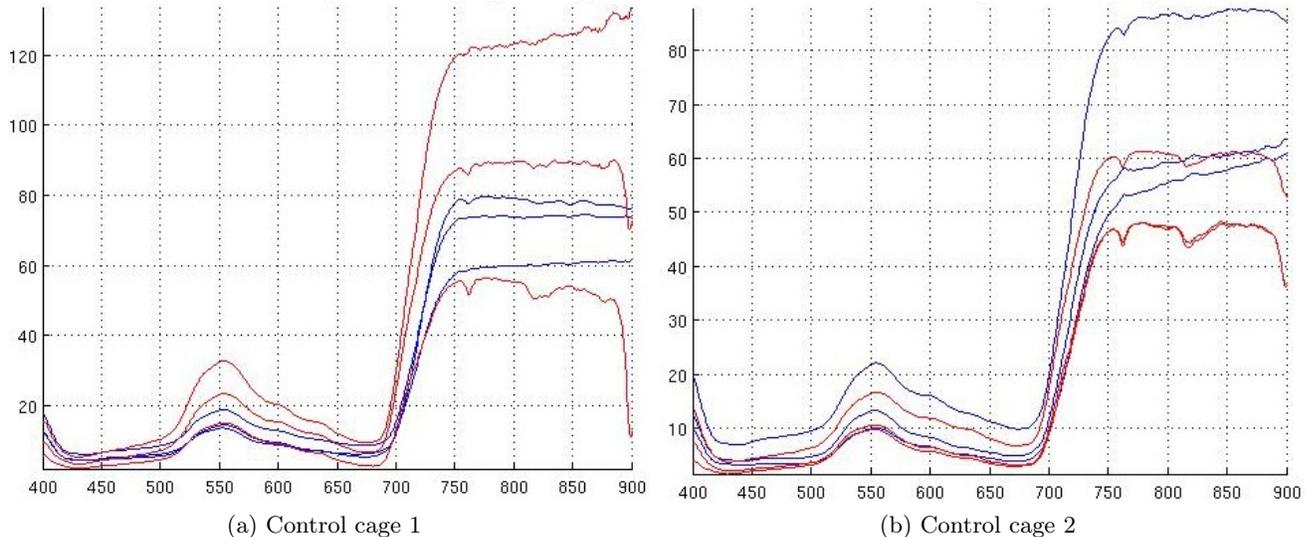


(a) Control cage 1
(b) Control cage 2

Figure 5: *Control cages.* The blue lines are the spectral response of three leaves extracted in the first campaign. The red lines are the spectral response of three leaves acquired nine days later. See text for details.

Note that the plants stress increased in all four plants. This is certainly due to the presences of first stage caterpillars that were certainly present within the cages. In any case, it can readily be seen that stress signs are significantly stronger in *Caterpillar cage 1* and *Caterpillar cage 2*, and that the trend in the features considered in Section 4 are consistent with our previous observations.

## 7. CONCLUSIONS AND FUTURE WORK

In this work we have demonstrated that automatic detection of plants suffering from stress induced by the presence of caterpillar is possible, by means of hyperspectral mesurements. Using a field spectrometer, we achieved a classification error as low as 95%. These results are extremely encouraging, since they open the path to effective remote sensing hypespectral imaging to distinguish healthy plants from those infected by caterpillar. The use of drones is actually a possibility, since a spatial resolution of 2 cm per pixel can be easily achieved. Moreover, it is important to note that once the relevant bands for the extracted features were detected, a solution involving the use of multispectral cameras could be possible.

Another interesting conclusion is the fact that, although not surprising, the stress induced by caterpillar is a property of the plants, and not just of each defoliated leave. Measurements on leaves of the same plant with different degrees of defoliation yielded the same spectral response.

In the near future, we plan to develop a methodology to classify different kinds of stress, mainly into biotic and abiotic. We already have some preliminary results that are very promising. We also plan to follow the path
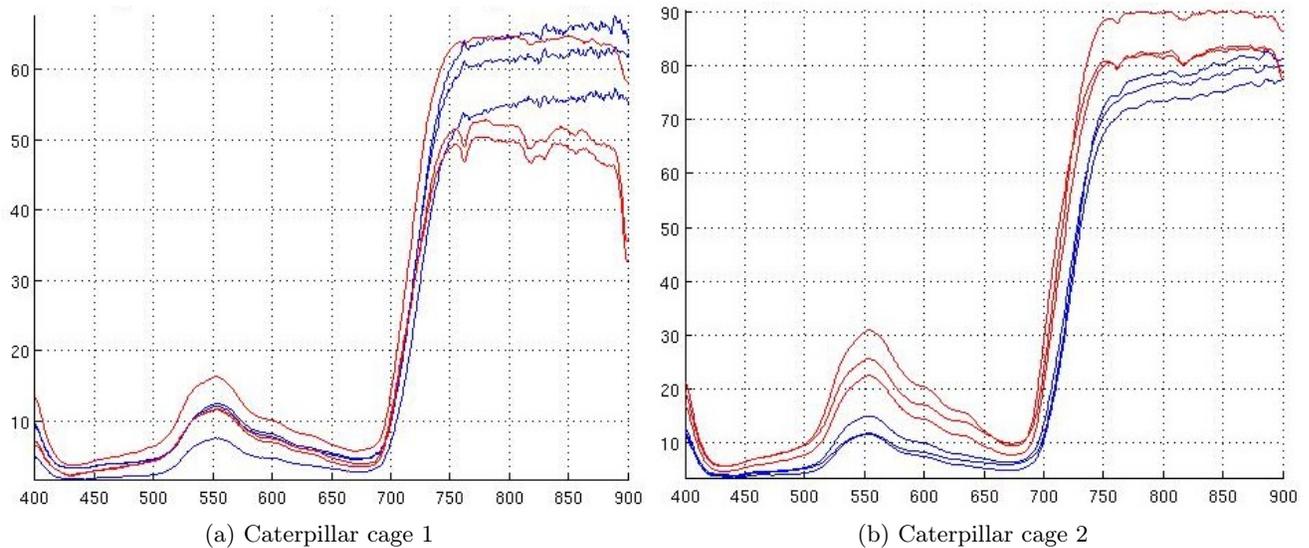
(a) Caterpillar cage 1             (b) Caterpillar cage 2

Figure 6: *Caterpillar cages.* The blue lines are the spectral response of three leaves extracted in the first campaign. The red lines are the spectral response of three leaves acquired nine days later. See text for details

of remote sensing detection of soy plants diseases, now that we have shown by means of fields spectrometry that such a classification is indeed possible.

## REFERENCES

[1] Jones, H. G. and Vaughan, R. A., eds., [*Remote sensing of vegetation*], Oxford University Press (2010).

[2] Thenkabail, P. S., Lyon, J. G., and Huete, A., eds., [*Hyperspectral Remote Sensing of Vegetation*], CRC Press (2011).

[3] McKinion, J. M., Jenkins, J., Willers, J., and Zumanis, A., "Spatially variable insecticide applications for early season control of cotton insect pests," *Computers and Electronics in Agriculture* **67**, 71–79 (2009).

[4] Yang, Z., M.N.Rao, Elliot, N., Kindler, S., and Popham, T., "Differentiating stress induced by greenbugs and russian wheat aphids in wheat using remote sensing," *Computers and Electronics in Agriculture* **67**, 64–70 (2009).

[5] Backoulou, G. F., Elliot, N., Giles, K., Phoofolo, M., and Catana, V., "Development of a method using multispectral imagery and spatial pattern metrics to quantify stress to wheat fields caused by diuraphis noxia," *Computers and Electronics in Agriculture* **75**, 64–70 (2011).

[6] Backoulou, G. F., Elliot, N., Giles, K., and Rao, M., "Differentiating stress to wheat fields induced by diuraphis noxia from other stress causing factors," *Computers and Electronics in Agriculture* **90**, 47–53 (2013).

[7] Zhang, M., Qin, Z., Liu, X., and Ustin, S., "Spectral discrimination of phytophtora infestants infection on tomatoes base don principal components and cluster analyses," *International Journal of Remote Sensing* **26**, 1095–1107 (2002).

[8] Zhang, M., Liu, X., and Ustin, S., "Detection of stress in tomatoes induced by late blight disease in california, usa, using hyperspectral remote sensing," *International Journal of Applied Earth Observation and Geoinformation* **4**, 295–310 (2003).

[9] Thenkabail, P. S., Smith, R. B., and Pauw, E. D., "Hyperspectral vegetation indices and their relationship with agricultural crop characteristics," *Remote Sensing and Evironment* **71**, 158–182 (2000).

[10] Thenkabail, P. S., Enclona, E. A., Ashton, M. S., and Meer, B. V. D., "Accuracy assessments of hyperspectral waveband performance for vegetation analysis applications," *Remote Sensing of Environment* **91**(3-4), 354–376 (2004).

[11] Bishop, C. M., [*Pattern Recognition and Machine Learning*], Springer (2006).

[12] Hastie, T., Tibshirani, R., and Friedman, J., [*The Elements of Statistical Learning*], Springer Series in Statistics, Springer New York Inc., New York, NY, USA (2001).

[13] Scholkopf, B. and Smola, A. J., [*Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*], MIT Press, Cambridge, MA, USA (2001).

[14] Witten, I. H., Frank, E., and Hall, M., [*Data Mining: Practical Machine Learning Tools and Techniques*], Morgan Kaufmann (2011).